# Beyond Translation Memories

**Reinhard Schäler**

Localisation Research Centre (LRC)
Department of Computer Science and Information Systems (CSIS)
University of Limerick
Limerick
Ireland
Reinhard.Schaler@ul.ie

**Abstract**

One key to the success of EBMT is the removal of the boundaries limiting the potential of translation memories. To bring EBMT to fruition, researchers and developers have to go beyond the self-imposed limitations of what is now traditional, in computing terms almost old fashioned, TM technology. Experiments have shown that the probability of finding exact matches at phrase level is higher than the probability of finding exact matches at the current TM segment level. We outline our implementation of a linguistically enhanced translation memory system (or *Phrasal Lexicon*) implementing phrasal matching. This system takes advantage of the huge and underused resources available in existing translation memories and develops a traditional TM into a sophisticated example-based machine translation engine which when integrated into a hybrid MT solution can yield significant improvements in translation quality.

## Translation Memories and EBMT

Having kept translators up at night worried about the future of their profession, machine translation as we knew it can now safely be declared obsolete and dead. However, since this initial threat to well established work practices, translation has never been the same. New developments in computer assisted translation, more specifically the emergence of translation memory (TM) applications, have continued to keep translators on their guard.

### TMs – sophisticated search & replace engines?

Initially, computational linguists often chose to simply ignore TM technology – considering it as some type of sophisticated search and replace engine, not a subject for serious research efforts.

The developers of these systems on the other hand, many of them coming from an academic background and therefore being familiar with the latest computational linguistic research developments, recognised the value of existing research and decided that it was time to apply it in practice.

For example: bilingual text alignment – this problem was declared as largely 'solved' in the early 90s by Gale and Church (1991) but never applied and proven in practice until the alignment utilities of translation memory developers some years later.[1]

Only recently, and driven by increased activities in the area of example-based machine translation (EBMT), has the interest shown by the linguistic tools industry in research results been reciprocated by the research community.

One possible reason for this development is that although EBMT as a paradigm has been described in research papers as far back as 1984 (Nagao etc.) and although it managed to capture the interest and enthusiasm of many researchers it has, so far, failed to reach the level of maturity where it could be transformed from a research topic into a technology used to build a new generation of machine translation engines – and new approaches, technologies and applications are badly needed in MT.

### Unlocking the potential of TMs

We believe that the time is ripe for the transformation of EBMT into demonstrators, technologies and eventually commercially viable machine translation engines along the lines suggested by Schäler (1996) and Macklovitch (2000) which are both based on the believe that existing translations contain more solutions to more translation problems than any other available resource (Isabelle et al., 1993).

The key to the success of this development, we suggest, is the removal of the boundaries limiting the potential of translation memories. To bring EBMT to fruition, researchers and developers have to go beyond the self-imposed limitations of what is now traditional, in computing terms almost old fashioned, TM technology.

## EBMT and the Phrasal Lexicon

EBMT has been proposed as an alternative and replacement for RBMT, initially by (Nagao, 1984), followed by extensions reported in (Sato & Nagao, 1990) and (Sadler & Vendelmans, 1990). EBMT has also been proposed as a solution to specific translation problems, as reported in (Sumita & Iida, 1991).

The enormous variety of approaches to, the focus of, and the motivations for the use of examples in natural language processing (NLP) are testimony to the high level of interest in EBMT. Taking existing parallel texts as their starting point, researchers have worked on:[2]

- Word-sense disambiguation (Brown et al., 1991) and translation ambiguity resolution (Doi, 1992) and (Uramoto, 1994);

[1] It is interesting to note here that what was deemed to be solved in theory turned out to present quite considerable problems when applied in practice (Schäler, 1994)

[2] The work quoted in the following bullet list can, unfortunately, not be fully referenced in this article, for practical reasons. Most of the reports mentioned were published in the proceedings of ANLP, COLING and ACL.

- Lexicography, e.g. the identification and translation of technical terminology (Dagan and Church, 1994), the development of an instant lexicographer (Karlgren, 1994); generally, the acquisition of lexical knowledge through the structural matching of bilingual sentences (Utsuro et al., 1994);
- Extraction of bilingual collocations or translation patterns from parallel corpora, non-aligned as in (Fung, 1995), (Rapp, 1995) and (Tanaka and Iwasaki, 1996), or aligned and using a linguistic (Matsumoto et al., 1993), statistical (Kupiec, 1993; Smadja et al., 1991; Smadja, 1993; Smadja et al.1996), or, indeed, a combined approach (Kumano and Hirakawa, 1994); special attention to the problems of extracting bilingual collocations for Asian languages is given by (Haruno, 1996) and (Shin, 1996);
- Translation Quality Measures (Su et al., 1992);
- Extensions to and variations of the basic idea of EBMT, proposing Pattern-based Machine Translation (Maruyama, 1993) and (Takeda, 1996), Transfer-Driven Machine Translation (Furuse and Iida, 1994), Statistical Machine Translation (Brown et al. 1993), Machine Translation based on Translation Templates (Kaji et al., 1992) and (Kinoshita, 1994), and Translation Patterns (Watanabe 1993) and (Watanabe, 1994).

One idea, however, which precedes all of the approaches mentioned and which, surprisingly, has so far not been taken up by researchers to any significant degree, that of the Phrasal Lexicon, described by Joseph Becker (1975).

## The Phrasal Lexicon

*"Like all other scientists, linguists wish they were physicists. They dream (...) of having language behave in an orderly way so that they could discover the Universal Laws behind it all. Linguists have a problem because language just ain't like that."* (Becker, 1975:70)

### Phrases as building blocks

Becker proposes a model radically different from that of mainstream linguistics as it is known since the mid-seventies: instead of considering language production as the process of combining units the size of words or smaller to form utterances, he heads in the opposite direction identifying phrases consisting of more than one word as the building blocks for the formation of utterances.

His thesis is that humans produce language mostly by repetition, modification and concatenation of previously-known phrases which are adapted to new situations during the productive process. While he concedes that *generative* processes ("generative gap filling") play an important role in language production, he believes that there is sufficient evidence to suggest that the use of language is at least as much based on memorization as on the generation of novel utterances because many situations do not demand novelty, but "rather an appropriate combination of formulas, clichés, idioms, allusions, slogans, and so forth".

**Becker's Theory of Language Production**

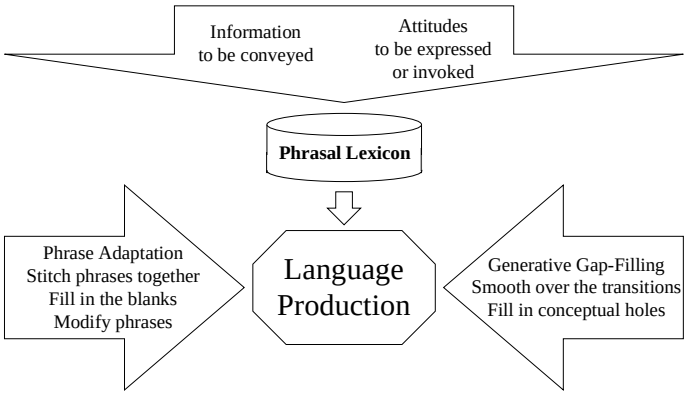


Figure 1: Becker's Theory of Language Production

For Becker, language generation is compositional[3] in the way illustrated in figure 1: the phrasal lexicon provides patterns that can provide (at least) some of the expressions needed to convey a message in a certain 'tone'. These phrases are then stitched together, blanks filled in and phrases modified where necessary. If this is not sufficient to generate the utterance, new phrases are generated to smooth over transitions and fill in conceptual holes.

### Recover observed linguistic behaviour

Becker's main motivation is to recover the observed linguistic behaviour of the native speaker as a subject for linguistic research. He feels that modern linguists who are working with artificially constructed models - namely the competence ("what language would be like if a decent mathematician had drafted it") and the performance ("everything that people actually say, write or think") models - make the abstract competence language model the subject of their research and deny the existence of language, e.g. English, *as she is spoke* for the purpose of communication between ordinary human beings.

#### Purely mathematical approach not valid

Becker hits hard against linguists who "dream of performing classic feats like dropping grapefruits off the Leaning Tower of Pisa (...) and in general of having language behave in an orderly way so that they could discover the Universal Laws behind it all". He believes that a purely mathematical approach to language is not only scientifically dishonest - because it disregards nearly all of its subject matter - but will ultimately be self-defeating yielding the scientific initiative to "untutored computer types" trained to "confront large-scale, complex, inelegant, real-world behavioural systems such as language, and attempt to understand the workings of

[3] Becker's notion of 'compositionality' is different from that used in the context of generative linguistics or, in fact, that used in rule-based MT (RBMT), i.e. that a target structure can be *generated* in a compositional manner (combining units the size of words or smaller) following a detailed analysis of the source structure and the establishment of correspondences between grammatical descriptions of the source and the target structures.

the systems without vainly pretending that they can be reduced to pristine-pure mathematic formulations."
According to Becker, dealing with lexical phrases has the additional advantage over transformations "and other such chimeras" in that they are actually observable - because they are real. [4]

## TM resources underused

In TM systems, two intellectually challenging problems have to be addressed which cannot just be solved by clever engineering.

**Fuzzy matches**
One is the decision of how to deal with cases where no exact matches can be found. Developers generally opt to search for similar matches and to calculate a ranking of identified 'fuzzy' matches which are then offered to the translator as a possible base for the translation of the new segment.

**Initial TM (alignment)**
The other problem occurs when translators want to create *initial translation memories* by aligning previous translations with their source on a segment-by-segment basis in order to import these aligned segments into a TM and then use this for the translation of a new version of the same source. Developers generally offer alignment tools which work in either interactive or fully automatic mode.

### Limitations

While the former problem still remains to be solved – in theory and in practice – the latter has been solved to a large extend. Remaining problems are purely engineering problems.
The availability of alignment tools linked with the now wide-spread use of translation memory systems has lead to the creation of massive bi- and multilingual parallel corpora aligned at sentence level[5] – the only segment level currently accessible by TM systems.
However, matching segments at sentence level unnecessarily restricts the potential and the usefulness of translation memories as extremely valuable linguistic resources for the following reason.
Returning to the first problem described as intellectually challenging, i.e. fuzzy matching: if a TM system cannot find an exact match in a TM, it can only propose fuzzy matches, i.e. matches where parts of the old and new source overlap leaving the task of adapting the translation of the old source to the new source up to the translator.

[4] It should be noted that Becker's categories for phrases are different from those generally accepted by English grammarians where each phrase is named after the word class of the head of the phrase, i.e. noun phrase, verb phrase, adjective phrase, adverb phrase, and prepositional phrase. (Greenbaum, 1996:208)
[5] Segments currently used by TM systems can be defined to a certain degree by users of TM systems and can also include text strings defined in documents such as headers or members of lists. Segments, however, can never be defined using linguistic criteria.

This can be a highly complex operation and, in fact, so cumbersome that translators often opt out of the fuzzy match proposal operation by setting the percentage threshold of the fuzzy match component so high that high percentage matches which could contain matching phrases are hidden away from them. Instead, they prefer to translate the new source without the support of the TM system to save time – valuable matches at sub-sentence level are lost.
The probability of finding exact matches at a lower phrasal level (e.g. at NP, VP or PP level) is significantly higher than the probability of finding exact matches at the current sentence level (i.e. the current TM segment level) as experiments have shown.
Storing, matching and proposing segments at phrasal level has a number of advantages, among them:

- Translators will be offered a higher percentage of exact matches from translation memories.
- The use of information stored in translation memories will increase – matching phrases in otherwise fuzzy matching sentences will no longer fall below the match percentage threshold set by most translators.
- The quality of translations produced by translators using translation memories will increase.
- TM systems will be able to translate a larger amounts of source text automatically without the need to adapt fuzzy matches manually.

Translation memories implementing phrasal matching will loose much of their appeal as intelligent but basically simple search and replace engines and become sophisticated example-based machine translation engines which when integrated into a hybrid MT solution will yield significant improvements in translation quality.

## TM and PL in EBMT

As outlined earlier, one approach to extending the linguistic coverage of TM systems is the redefinition of a translation unit or segment in a situation where only a fuzzy match can be identified. In such a case, translation units could be defined at phrase level. These phrasal units could then be looked up in a phrasal lexicon and be translated by combining already translated phrases stored in the phrasal lexicon – very much along the lines proposed originally by Becker.
Using a TM containing the two entries:

| |
|---|
| [1]<br>[ENG] The bullets move to the new paragraph.<br>[GER] Die Blickfangpunkte rücken in den neuen Abschnitt.<br>[2]<br>[ENG] The title moves to the centre of the slide.<br>[GER] Der Titel rückt in die Mitte des Dias. |

Table 1: TM entries

The following new sentence could not be translated automatically:

| The bullets move to the centre of the slide. |
|---|

Table 2: New source sentence

At most, the system would be capable of identifying one of the two sentences in the TM as a fuzzy match and display its translation which would then have to be adapted by a translator.

## Higher match values at a price

If the translation memory, however, could provide translation units at phrase level those units could be looked up in a *phrasal lexicon* and combined so that the system could produce a correct translation of the new sentence.

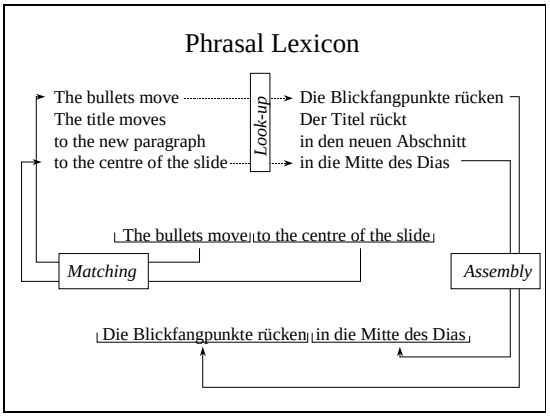


Figure 2: The Phrasal Lexicon - Overview

The phrasal lexicon has the potential to offer translators many of the advantages of a translation memory (consistency, cost savings - no coding or post-editing as with MT systems). However, because it is based on smaller translation units it can potentially identify more exact matches than the sentence based TM systems - but at a price.

**Linguistic processing**

TM systems do not have to carry out any significant amount of linguistic processing, e.g. they practically do not need to know anything about the target language as most of the processing (matching, calculation of fuzziness, identification of changes etc.) is being done with the source language.

By contrast, a PL would need to know - at a very minimum - enough about the source *and* the target language to identify phrases and describe the linguistic characteristics of their constituent parts. This includes the need to parse source and target language sentences and the ability of the parser used by the system to select one parse for inclusion into the PL while being able to deal with the issues inherent in such a process, including ambiguities, failing parses and wrong parses.

At the same time, the amount of linguistic processing expected to be performed by a PL systems and the human effort in maintaining it will still be significantly lower than in the case of a full MT system.

Therefore, in the context of automated translation tools, PL systems can be positioned in between MT and TM, ideally suited to deal with 'familiar' text (or 'fuzzy matches' in TM terminology) while MT would continue to deal with unfamiliar text (*no match*) and TM with unchanged text (*100% match* at sentence or paragraph level) as illustrated in Figure 3.

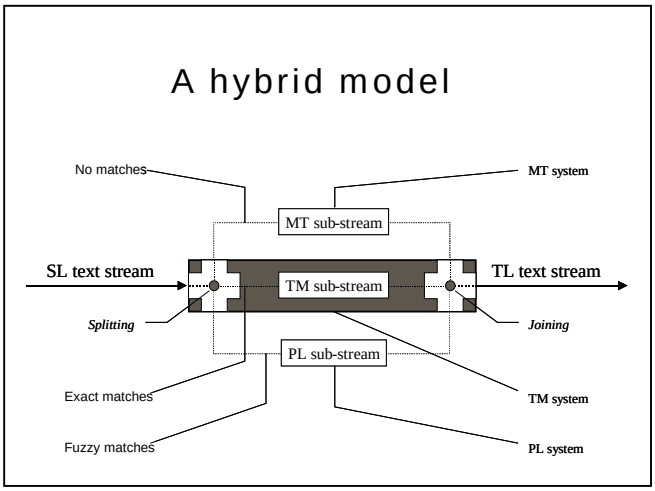


Figure 3: The Phrasal Lexicon (PL) as part of a hybrid MT environment

Frederking and Nirenburg (1994) were probably the first to propose a multi-engine MT architecture, passing the input to multiple translation technologies in parallel and combining their outputs to generate a final translation.

Using 'traditional' MT in combination with the now also established TM approach, together with the new PL engine is one implementation model for this approach.

## Phrasal Matching

Different approaches are possible to implement matching at phrase level. The following paragraphs describe our approach taken for the implementation of a demonstrator, the Phrasal Lexicon (PL).

The underlying linguistic technology for this approach was developed by Allan Ramsay whose original parser and grammar for English were extended to cover both English and German with a separate dictionary for each of the languages (Ramsay & Schäler, 1995).

The demonstrator implements a number of important functions:

- The use of one parser and one grammar for English and German to analyse a source and a target text.
- The automatic or interactive/semi-automatic production of a bilingual phrasal lexicon (including linguistic annotation).
- The "translation" of a new source text by matching phrasal units from the PL (by combining of substituting known phrases).

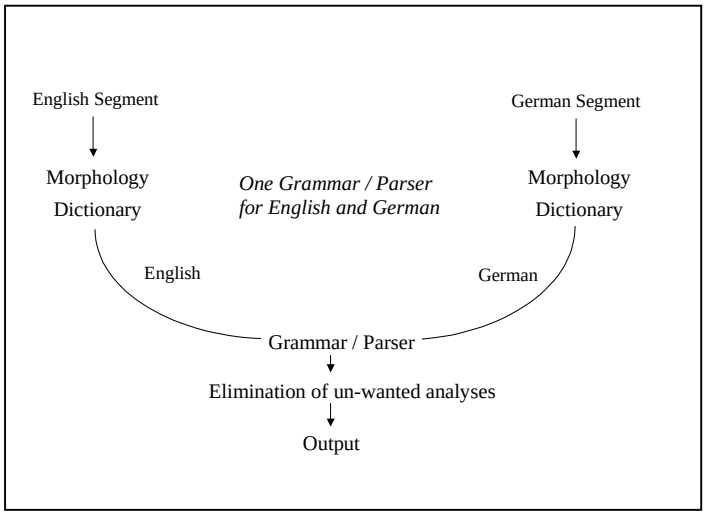


Figure 4: Parsing for the PL

## Building a PL

Source language can be provided to the PL in two ways.

- As a file containing bilingual sentence pairs, e.g. exported translation memories.
- Interactively by a translator typing in source and target sentences.

Sentences are read and then analysed by the system's grammar and parser. Source and target language segments are segmented into phrasal units which are stored in a phrasal lexicon together with relevant linguistic information.

As mentioned earlier, only one parser and one grammar are used for the analysis of source and target language segments which ensures comparable output for each of the languages. The only difference is in relation to the dictionary of where there is one for each language.

This system is able to automatically process previously created translation memories without human intervention and transform it into a *linguistically enhanced* translation memory, a phrasal lexicon, thus transgressing the boundaries of traditional translation memory technology.

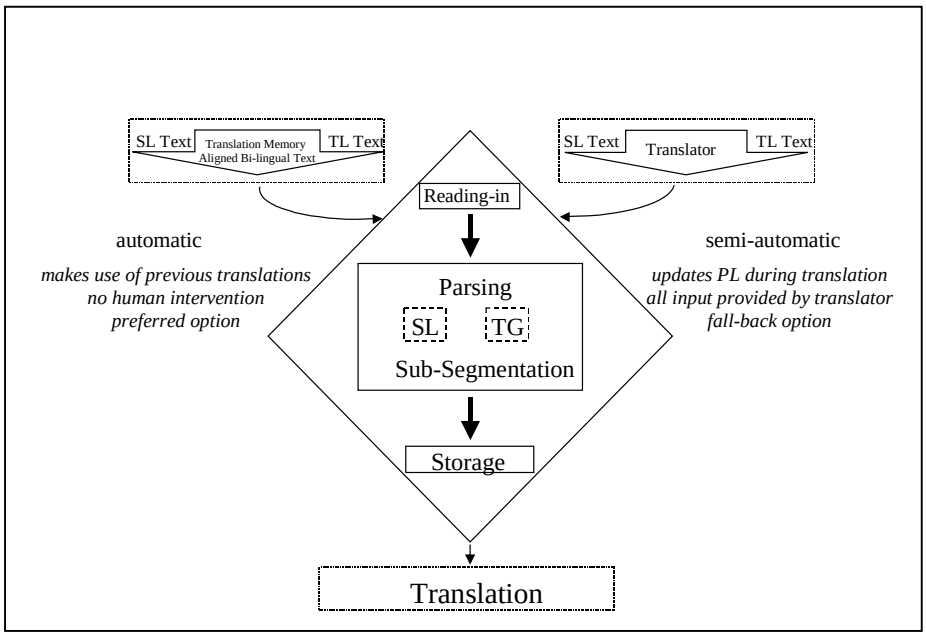


Figure 5: Building a Phrasal Lexicon

Entries in a phrasal lexicon contain multilingual phrase pairs, together with a description of some of their linguistic characteristics.

```
phrase_pair
(sign (phonology (structure
       ([he,{buy,ed},a,{car,''}],_,_,_),_,_,
              phon_type(_,_,_,_,english),_,_),
              syntax(nonfoot(major(cat(n(-),v(+)),bar(_,phrasal(+))),
              head(agree(_,_,third(sing(+),plural(-)),_,_),
              vform(vfeatures(finite
                     (tense(present(),past(+)),
                     tensed(+),participle(-),infinitive(),
                     to_form(-)),_,_,_,_,_,_),_),_,_),_,_),_),
sign(phonology(structure
       ([er,{kauf,te},ein,{'Auto',''}],_,_,_),_,_,
              phon_type(_,_,_,_,german),_,_),
              syntax(nonfoot(major(cat(n(-),v(+)),bar(_,phrasal(+))),
              head(agree(first(_,plural(-)),second(sing(-),plural(-
                     ),third(sing(+),plural(-)),count(individual(+),kind(-
                     ),mass(-),_),gender(neuter(-),male(+),female(-))),
              vform(vfeatures(finite
                     (tense(present(),past(+)),
                     tensed(+),participle(-),infinitive(-),
                     to_form(-)),_,_,_,_,_,_),_),_),_,_),_,_),_)).
```

Figure 6: PL sample entry

## Translating new sentences

In the hybrid system described earlier, only sentences described as *familiar* (fuzzy matches in TM terminology) will be submitted to the PL component. These sentences are then parsed and analysed.

The PL checks if there is a corresponding syntactical structure in the PL. If that is the case the system matches the surface forms of both the source and the target language phrasal units using a bi-lingual index to certify the phrasal units which are then used to generate a new target language segment.

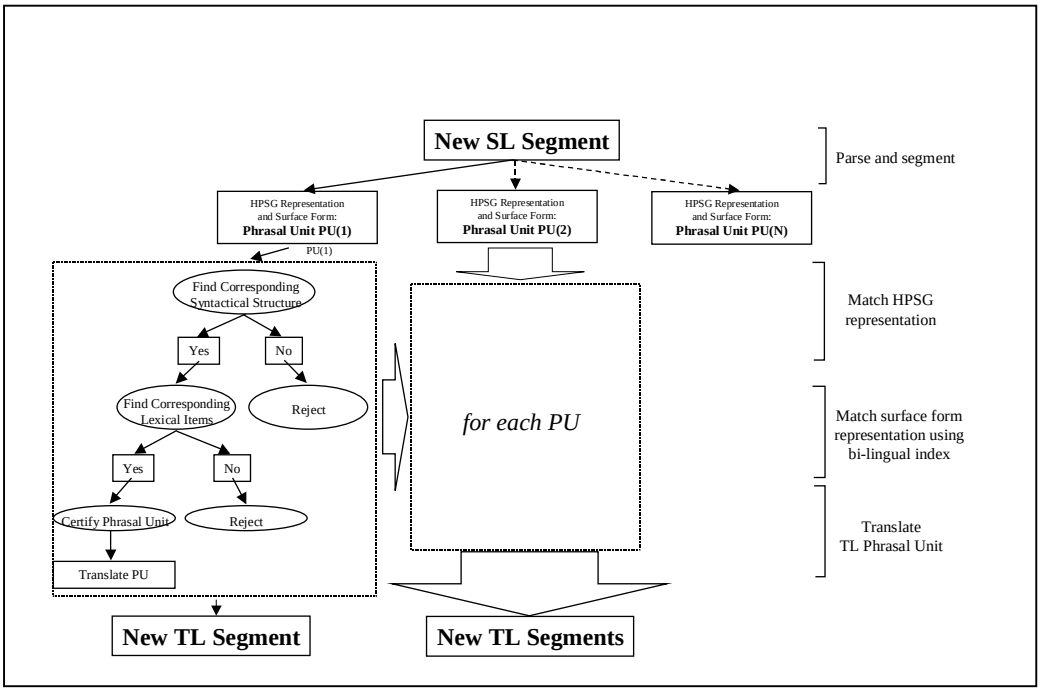


Figure 7: PL translation component

## Conclusion

For many decades and on various occasions, MT researchers have claimed that the linguistic technology developed by them has made manual, human translation redundant. These claims have so far not had a significant impact on the reality of translation as a profession and as a business.

The one technology that did have an impact on translation has been TM – it changed the way translators work as can be seen when examining the impact it had in the localisation industry, one of the largest employers of technical translators.

Ironically, TM technology worked without any of the sophisticated linguistic technologies developed over decades by MT developers – it is little more than a sophisticated search and replace engine.

However, because of the enormous success of TM systems, there are now large amounts of TMs available – exactly how many can only be estimated: individual products, which are frequently translated into 30 languages and more, can easily contain up to one million words.

But the highly successful approach taken by TM developers is also the cause for the inherent restrictions and limitations of TMs.

To overcome these, we have proposed an implementation of EBMT based on the idea of a phrasal lexicon (or a linguistically enhanced version of a TM system working at phrase level). A first demonstrator of this system has been built and is currently being evaluated.

## Acknowledgements

Initial research on this project was conducted at University College Dublin (UCD) under Prof. Allan Ramsay, now at UMIST. It was part-funded by the Irish Industrial Development Agency (IDA).

## Bibliographical References

Becker (1975). Joseph D. Becker, Bolt Beranek and Newman. The Phrasal Lexicon. In Proceedings of Theoretical Issues in Natural Language Processing, pages 70-73. Cambridge, Massachusetts (10-13 June 1975).

Frederking and Nirenburg (1994). Three Heads are better than one. R Frederking and S. Nirenburg. In Proceedings of Applied Natural Language Processing (ANLP), 1994.

Gale, William A. & Church, Kenneth W. (1991). A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, June 1991.

Isabelle, P. et al. (1993). Translation Analysis and Translation Automation. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, 1993.

Macklovitch, Elliot (2000). Two types of Translation Memory. In Translating and the Computer 22, Proceedings of the ASLIB Conference, London (16-17 November 2000).

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji, editors, Artificial and Human Intelligence, pages 173-180, North Holland, Amsterdam, 1984.

Planas, E. & Furuse O. (1999). Formalizing Translation Memories. In Proceedings of MT Summit VII, Singapore (13-17 September 1999), pp 331-339.

Ramsay, Allan & Schäler, Reinhard (1995). *Case and Word Order in English and German*, Allan Ramsay and Reinhard Schäler, in: Ruslan Mitkov and Nicolas Nicolov (eds.): *Recent Advances in Natural Language Processing*, John Benjamins (Amsterdam/Philadelphia) 1997.

Sadler (1990). Victor Sadler and Ronald Vendelmans. Pilot Implementation of a Bilingual Knowledge Bank. Proceedings of the Conference on Computational Linguistics (COLING) 1990, pp. 449-451.

Sato, S. & Nagao,M. (1990). Toward Memory-based Translation. In Proceedings of the Conference on Computational Linguistics (COLING) 1990, pp 247-252.

Schäler, Reinhard (1994). A Practical Evaluation of an Integrated Translation Tool during a Large Scale Localisaton Project. In Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, Germany (October 13-15, 1994).

Schäler, Reinhard (1996). Machine translation, translation memories and the phrasal lexicon: the localisation perspective. In TKE 96, EAMT Machine Translation Workshop, Vienna, Austria (29-39 August 1996), pp. 21-33.

Sumita, E. & Iida, H. (1991). Experiments and Prospects of Example-based Machine Translation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), 1991, pp 185-192.